FINAL REPORT ON PLAN FOR THE ASSESSMENT AND EVALUATION OF INDIVIDUAL AND TEAM PROFICIENCIES DEVELOPED BY THE DARWARS ENVIRONMENTS

Harold F. O'Neil

Eva L. Baker

Richard Wainess

Claire Chen

Robert Mislevy

Patrick Kyllonen

Advance Design Information

Sherman Oaks, CA

December 31, 2004 v.5

The work reported herein was supported in part under Office of Naval Research Award Number #N00014-03-C-0357, as administered by the Office of Naval Research, and in part by Defense Advanced Research Projects Agency. The findings and opinions expressed in this report do not reflect the positions or policies of the Office of Naval Research or Defense Advanced Research Projects Agency.

Contents

Executive Summary

The goal of this final report is to produce a first cut for a feasible, valid, and cost-sensitive evaluation plan to permit the development and application of metrics to assess the impact of participating in a DARWARS environment. DARWARS is a DARPA program whose goals have evolved over time due to changing requirements and technology opportunities. The original goal as reflected in the RFP was to transform military training by providing continuously available, on-demand, mission-level training for all forces at all echelons. DARPA proposed to create an environment where there will be a set of continuously available virtual training wars in which any unit or individual can participate via personal computer-based systems that will teach and exercise users in the cognitive and affective aspects of their warfare areas via electronic wargames. The current goal as reflected in documentation at the 2004 Interservice/Industry Training, Simulation and Education Conference is "to accelerate the development and deployment of the next generation of experiential training systems. These low-cost, web-centric, simulation-based systems take advantage of the ubiquity of the PC and of new technologies, including multiplayer games, virtual worlds, intelligent agents, and on-line communities. DARWARS training systems offer engaging practice environments for individuals and teams with on-target feedback for each student" (BBN/DARWARS, 2004, p. 2, col. 1).

The goals of our planning effort were to operationalize evaluation questions, identify existing measures and metrics that could be employed, communicate with the providers of DARWARS software to assure that an evaluation will address the key intended outcomes, identify unintended outcomes, and propose a way to integrate findings across different implementations. In addition, assistance to individual projects was also provided. The date of

award of our contract was July 11, 2003. Two reports were prepared: a draft report (September 30, 2003) and this, our final report.

While effectiveness of game environments can be documented in terms of intensity and longevity of engagement (participants voting with their quarters or time), as well as the commercial success of the games, there is little solid information about what outcomes are systematically achieved by the use of multiplayer games to train participants in acquiring knowledge and skills. What is missing is how games should be evaluated—for example, the degree to which games are designed to foster the key skills and strategies desired in both individuals and teams. Evaluation questions to be answered about the cognitive and affective effects of games should concern the four levels of Kirkpatrick's (1994) framework. Another form of impact involves the development of social capital among the players. For example, questions should be raised about the social capital that the game environment produces, and in particular about teamwork, deep knowledge of partners and opponents, and how a sense of collective efficacy and trust develops.

In this document, we provide a series of theoretical assessment and evaluation frameworks. We focus on both the cognitive and affective effects of games. A literature review, an evaluation framework, our technical approach to evaluation, availability of editors, and a description of an experimental study assessing the training effectiveness of a game are provided.

Statement of Problem

While effectiveness of game environments can be documented in terms of intensity and longevity of engagement (participants voting with their quarters or time), as well as the commercial success of the games, there is much less solid information about what outcomes are systematically achieved by the use of individual and multiplayer games to train participants in acquiring knowledge and skills. For example, early studies of the game "WEST" compared variations in coaching but found that the game itself produced little increment in learning (Baker, Aschbacher, & Bradley, 1985). Thus, the game was not instructionally more effective than traditional approaches. Similar results were found by Parchman, Ellis, Christinaz, and Vogel (2000). Yet, there is a persistent belief that games can provide an attractive venue for engaging participants in learning. They can produce rapid multiple trials, allow students to manage their feedback or to replay with a different strategy, and include social components, such as teamwork, of real value in the military.

What is missing is how games should be evaluated for training purposes. First is the degree to which they are designed to foster key knowledge and skills desired. Secondly, the impact of game playing needs to be studied to determine what works (see O'Neil, Baker, & Fisher, 2002; O'Neil & Fisher, 2004). Without an investment in evaluation and the accumulation of clear evidence of impact, there will be a tendency to dismiss game environments as motivational fluff.

Not surprisingly, the evaluation of complex, multi-user games, in particular, presents a number of challenges—for example, team training and assessment. There is the problem of identifying obvious and non-obvious outcomes. For instance, it will make a difference in data

flow and cost to decide which criterion behaviors should be addressed and with what frequency, and the degree to which measures can be integrated seamlessly into the games.

There are multiple ways of assessing learning. For example, one could specify the assessment of the training effects of a game by examining trainees' ability to solve criterion problems, their application of declarative and procedural knowledge, their willingness to raise or lower game challenge conditionally, their self-reports, and records of their play. Evaluation questions to be answered about the cognitive and affective effects of games should concern the four levels of Kirkpatrick's (1994) framework. The evaluation system should include measures related to the attainment at different levels of expertise of specific content and skill acquisition being trained. These may include skills to be learned along the way, as well as those of an outcome nature.

There is also a skill related to personal development—the "learning to learn" or self-regulation outcome. To what degree do participants develop the strategic knowledge necessary to apply to the specific training topics? Do players develop more general predispositions and behaviors that support transfer of knowledge across different contexts or problem variations? The development of individual metacognitive skills, especially as used in time-constrained environments, should be estimated. Robustness of advanced processes in learner planning, monitoring, and cognitive strategy selection could be assessed ideally in transfer situations in the light of the cost to develop such processes.

Another form of impact involves the development of social capital among the players. Questions should be raised about the social capital that the game environment produces, and in particular about teamwork, deep knowledge of partners and opponents, and how a sense of collective efficacy, effort, and trust develops. Teamwork is a good example: Not only does

cooperation yield benefit, but team members learn rapidly what expertise their teammates share. A measure of the team's mental model of competence is desirable, for instance, to see whether, over time, they are adapting to different skill, predisposition, or knowledge levels. These include monitoring the development of collective efficacy and effort. In addition, the collection of non-obtrusive information on human and team development would be essential.

At the outset, a clear vision should be articulated of the types of reports needed for audiences key to the continuation or expansion of the system. There must also be a way to translate the value of the evaluation itself, if there is a desire to build new systems with a continuous monitoring approach to detect growth and deficiencies. To support warfighter assessment in computational environments, one must fuse cognitive, conative, and behavioral data streams to model moment-to-moment warfighter states for both individuals and teams.

Maximizing skill and knowledge acquisition and retention in extremely short periods becomes tractable if there is comprehensive and accurate information on the trainee's background (e.g., quantitative and verbal aptitude, degree of prior knowledge, and experience in the training content), information on performance on the training task outcomes (e.g., quality of solution), and ongoing measures of behavioral (e.g., trainee's clickstream), conative (e.g., motivation, self-regulation), and cognitive processes embedded within the task (e.g., measures of trainee understanding, stress, cognitive load).

What Is DARWARS?

DARWARS is a DARPA program whose goals have evolved over time due to changing requirements and technology opportunities. The original goal as reflected in the RFP was to transform military training by providing continuously available, on-demand, mission-level training for all forces at all echelons. DARPA proposed to create an environment where there will

be a set of continuously available virtual training wars in which any unit or individual can participate via personal computer-based systems that will teach and exercise users in the cognitive and affective aspects of their warfare areas via electronic wargames. The current goal as reflected in documentation at the 2004 Interservice/Industry Training, Simulation and Education Conference is "to accelerate the development and deployment of the next generation of experiential training systems. These low-cost, web-centric, simulation-based systems take advantage of the ubiquity of the PC and of new technologies, including multiplayer games, virtual worlds, intelligent agents, and on-line communities. DARWARS training systems offer engaging practice environments for individuals and teams with on-target feedback for each student" (BBN/DARWARS, 2004, p. 2, col. 1).

Literature Review

Two literature searches were conducted using three search engines, PsycINFO, EducationAbs, and SocialSciAbs. The purpose of the first search was to locate articles that reported on research about the use of video games in general, for training adults. The second search was to locate articles that reported research specifying the use of multiplayer or massively multiplayer video games for training adults. The literature review is structured in the following manner: First two theoretical frameworks are provided and related to each other. Then we define critical terminology (i.e., games, simulation, and simulation games). Next, we provide the literature review results and a summary of research. Our prior report presented the methodology for the literature review. This final report will integrate the findings from the literature.

Educators and trainers began to take notice of the power and potential of computer games for education and training back in the 1970s and 1980s (Donchin, 1989; Malone, 1981; Malone & Lepper, 1987; Ramsberger, Hopwood, Hargan, & Underhill, 1983; Ruben, 1999; Thomas &

Macredie, 1994). Computer games were hypothesized to be potentially useful for instructional purposes and were also hypothesized to provide multiple benefits: (a) complex and diverse approaches to learning processes and outcomes; (b) interactivity; (c) ability to address cognitive as well as affective learning issues; and perhaps most importantly, (d) motivation for learning. We view the literature through two major theoretical frameworks: (a) Kirkpatrick's (1994) four levels for evaluating training and (b) Baker and Mayer's (1999) CRESST model of learning.

Kirkpatrick's Four Levels for Evaluating Training

Kirkpatrick (1994) described four levels that represent a sequence of ways to evaluate programs (Figure 1): Level 1 is *Reaction*, Level 2 is *Learning*, Level 3 is *Behavior*, and Level 4 is *Results*. According to Kirkpatrick, when moving from level to level, the evaluation process becomes more difficult and time consuming, but provides more valuable information. Kirkpatrick further argued that all levels should be included in an evaluation and should be assessed in order of the levels. *Reaction* (Level 1) is a measure of learner satisfaction and measures "how those who participate in the program react to it" (p. 21). Kirkpatrick commented that if learners do not react favorably to the training, they will probably not be motivated to learn. According to Kirkpatrick, "Positive reactions may not ensure learning, but negative reaction almost certainly reduces the possibility of its occurring" (p. 22). He described *Learning* (Level 2) as "the extent to which participants change attitudes, improve knowledge, and/or increase skill as a result of attending the program" (p. 22). Level 2 is evaluated within the context of the training session.

Four Levels for Evaluating Training	
Level 1: REACTION	Trainee's reaction to the program: Level of satisfaction.
Level 2: LEARNING	Trainee's attitude change, increased knowledge, and/or increase skill, due to the training.
Level 3: BEHAVIOR	On the job change in behavior because of program participation.
Level 4: RESULTS	How the organization benefited from the learner's participation in the program (e.g., increased production or profits, improved quality, decreased costs, less accidents).

Figure 1. Kirkpatrick's (1994) four levels for evaluating training.

Behavior (Level 3), which is evaluated on the job, after training has occurred, is defined as "the extent to which change in behavior has occurred because the participant attended the training program" (Kirkpatrick, 1994, pp. 22-23). Level 3 concerns the concept of *transfer*, which Brunken, Plass, and Leutner (2003) described as the ability to apply acquired knowledge and skills to new situations.

The final level, *Results* (Level 4), refers to the benefits from the company's perspective and can be defined as the "final results that occurred because the participant attended the program" (Kirkpatrick, 1994, p. 25). Final results are related to cost effectiveness of training or return on investment and include increased production, improved quality, decreased costs, reduced frequency and/or severity of accidents, increased sales, reduced turnover, and higher profits or return on investment. Generally these results are the reason for attending the program. The Kirkpatrick framework is the dominant one in training evaluation but is seldom used in the education sector.

Baker and Mayer's CRESST Model of Learning

The CRESST model of learning (Baker & Mayer, 1999) (see Figure 2) is composed of five families of cognitive demands: content understanding, collaboration or teamwork, problem solving, communication, and self-regulation. In the CRESST model, "each family consists of a task that can be used as a skeleton for the design of instruction and testing" (Baker & Mayer, 1999, p. 275). For example, content understanding involves explanation, which in turn involves a variety of actions such as having students read opposing views (of the content), invoking prior knowledge, and organizing and writing a valid explanation. Such knowledge and skills can be measured by multiple-choice tests, essays, or knowledge maps. This framework supports many different learning domains, such as history or science. For problem solving, there is a need to identify content understanding, problem-solving strategies and self-regulation skills (O'Neil 2002). Each of these aspects of problem solving would have then specific measures. Each of the other cognitive demands is also supported by specific constructs, instructional strategies, and measures (Baker & Mayer, 1999; Baker & O'Neil, 2002).

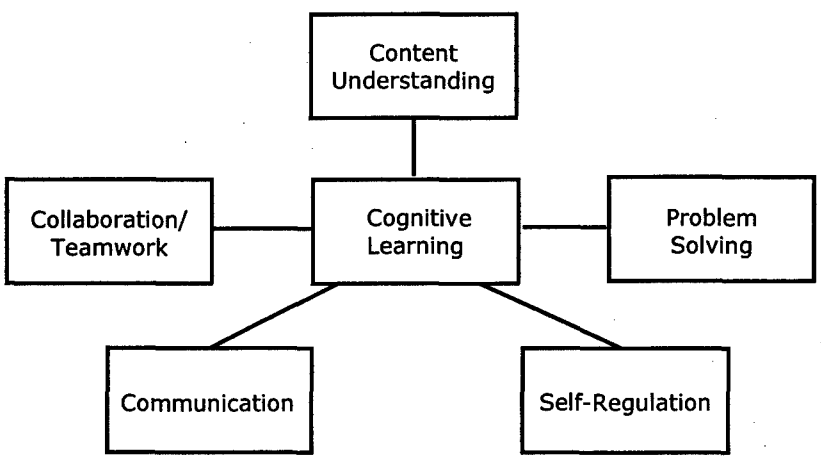

Figure 2. Baker and Mayer's (1999) CRESST model of learning: Families of cognitive demands.

Relating the frameworks. Kirkpatrick's (1994) four levels for evaluating training and Baker and Mayer's (1999) CRESST model of learning were not designed for the evaluation of research study outcomes per se. However, both frameworks were designed to evaluate learning. Since the goal of many studies, and in particular the studies in this review, is to examine characteristics that can affect learning outcomes, these two frameworks are particularly relevant and complementary. Kirkpatrick's is the macro view in evaluation, and the CRESST model of learning is the micro view in the area of learning.

There is no relationship between Kirkpatrick's (1994) Level 1 (Reaction) and CRESST's (Baker & Mayer, 1999) model of learning, as Kirkpatrick's Level 1 refers to satisfaction, not learning. Likewise, there is no relation in Baker and Mayer's model to Kirkpatrick's Level 4 (Results), since *results* refers to institutional benefits, not learner benefits per se. However, Kirkpatrick's Level 2 (Learning) equates to the entire CRESST model, and Kirkpatrick's Level 3 (Behavior) is reflected in transfer measures on the job, some of which would reflect the CRESST model of learning. We will use both frameworks in this report to review the literature in computer games.

Defining Games, Simulations, and Simulation Games

A major problem area with reviews of research on games and simulations is terminology. The three most commonly used terms in the literature are *game*, *simulation*, and *simulation game*, yet there is little consensus in the education and training literature on how these terms are defined. Because the goals and features of games and simulations (as well as the hybrid, simulation games) differ, it is important when examining the potential effects of the two media—games and simulations—to be clear about which one is being examined. Clearly defining the differences

between games and simulations also provides a foundation for accurately defining the mixed mode environment of simulation games.

This report will use the definitions of games, simulations, and simulation games as stated by Gredler (1996). These definitions combine the most common features cited by the various researchers, and yet provide clear distinctions between the three media. According to Gredler,

> Games consist of rules that describe allowable player moves, game constraints and privileges (such as ways of earning extra turns), and penalties for illegal (nonpermissable) actions. Further, the rules may be imaginative in that they need not relate to real-world events. (p. 523)

This definition is in contrast to that of a simulation, which Gredler (1996) defined as "a dynamic set of relationships among several variables that (1) change over time and (2) reflect authentic causal processes" (p. 523). In addition, Gredler described games as having a goal of winning whereas simulations have a goal of discovering causal relationships. Gredler also defined a mixed metaphor referred to as *simulation games* or *gaming simulations*, which is a blend of the features of the two interactive media, games and simulations.

In terms of goals, however, games and simulations differ. With games, once a goal is achieved, it cannot be repeated without intentionally reverting to a prior game state or restarting the game. That is, the flow of a game must be purposefully interrupted. In a game comprised of multiple goals, achievement of one goal results in commencement of work toward the next goal or set of goals. Therefore, with games, the goal structure is linear. In contrast, simulations have nonlinear goal structures. With simulations, the goal is to achieve a desired output state or simply to examine output states, based on the manipulation of input variables. Once the goal is achieved, the player can continually make modifications to the input variables, examining their effect on the output. Therefore, because this process can be repeated as often as desired, the goal structure of a simulation is nonlinear.

While this report uses Gredler's (1996) definitions, it should be noted that her definitions are not in agreement with those of other game researchers. Table 1 lists a summary of the characteristics that various researchers have attributed to games and to simulations. We have included our extension of Gredler's reference to linearity and nonlinearity by expressing linearity in terms of gameplay and goals separately. A column for simulation games was not included in the table since, as its name implies, simulation games would have combinations of game and simulation characteristics.

Table 1

Characteristics of Games and Simulations

Characteristic	Game	Simulation
Combination of ones actions plus at least one other's actions	Yes (via human or computer)	Yes
Rules	Defined by game designer/developer	Defined by system being replicated
Goals	To win	To discover cause-effect relationships
Requires strategies to achieve goals	Yes	Yes
Includes competition	Against computer or other players	No
Includes chance	Yes	Yes
Has consequences	Yes (e.g., win/lose)	Yes
System size	Whole	Whole or Part
Reality or fantasy	Both	Both
Situation specific	Yes	Yes
Represents a prohibitive environment (due to cost, danger, or logistics)	Yes	Yes
Represents authentic cause-effect relationships	No	Yes
Requires user to reach own conclusion	Yes	Yes
May not have definite end point	No	Yes
Contains constraints, privileges, and penalties (e.g. earn extra moves, lose turn)	Yes	No

Literature Review Results

A specific review to explore the classification of game literature using the frameworks of Kirkpatrick (1994) and the CRESST model of learning (Baker & Mayer, 1999) was conducted. Our literature review (last 15 years) using information systems (i.e., PsycINFO, EducationAbs, SocialSciAbs) and the search terms *games, computer game, PC game, computer video game, video game, cooperation game,* and *multi-player game* resulted in only 18 articles with either qualitative or qualitative information on the effectiveness of games with adults as participants. We selected only journal articles for review. Thus, research based on dissertations or technical reports was not examined. We then examined the viability of contextualizing games research using two frameworks; the Kirkpatrick (1994) levels and the CRESST model of learning (Baker & Mayer, 1999). We believe we can show support for the utility of these frameworks.

In the past 15 years, several thousand articles pertaining to games have been published. However, only 18 studies met our standards for empirical research and are reviewed in this report. As cited earlier, findings regarding the educational benefits of games are mixed, and it is hypothesized that the positive findings can be attributed to instructional design and not to games per se. Also discussed earlier was the issue that many studies claiming positive outcomes appear to be making unsupported claims for the media. These issues appear to be echoed in the studies we reviewed. For example, Mayer, Mautone, and Prothero (2002) examined performance outcomes using retention and transfer tests, and Carr and Groves (1998) examined performance outcomes using self-report surveys. Mayer et al. (2002) offered strong support for their findings, showing statistical significance for the treatment, whereas Carr and Groves used only participants' self-reports as evidence of learning effectiveness. In Carr and Groves' study, participants reported their belief that they learned something from the experience. No

performance was actually measured, yet Carr and Groves suggested that their simulation game was a useful educational tool, and that use of the tool provided a valuable learning experience.

Table 2 lists the media and measures utilized by the 18 studies to assess learning. We have also categorized the measures in Table 2 as to Kirkpatrick's (1994) levels and also type of learning in terms of Baker and Mayer's (1999) framework. Of the 18 studies, 9 used a single measure to examine outcomes. For example, Carr and Groves (1998) used self-report via a survey instrument, another study (Rosenorn & Kofoed, 1998) used observation, and a third used think-aloud protocols (Hong & Liu, 2003).

Table 2

Media, Measures, Kirkpatrick Levels and CRESST Categories

Study	Media[a]	Measures	Kirkpatrick levels[b]	CRESST categories
Arthur et al. (1995)	Space Fortress (a)	Performance on game	2	Collaboration, problem solving, communication, and self regulation
Carr & Groves (1998)	Business Simulation (c)	Survey	1, 2	Content understanding, collaboration, communication, and self regulation
Day, Arthur, & Gettman (2001)	Space Fortress (a)	Performance on game and knowledge map	2	Content understanding, problem solving, and self regulation
Galimberti, Ignazi, Vercesi, & Riva (2001)	3D-Maze (a)	Observation and time to complete game	1, 2	Collaboration, problem solving, communication, and self regulation
Gopher, Weil, & Bareket (1994)	Space Fortress (a)	Performance on game and transfer task	1, 2, 3, 4	Problem solving and self regulation
Green & Bavelier (2003)	Medal of Honor (c)	Visual attention measures	2	Content understanding
Hong & Liu (2003)	Klotski (a)	Think-aloud protocols	2	Problem solving
Mayer (2002)	Fifth Dimension (a)	Retention and transfer tasks	2	Problem solving, content understanding, self-regulation
Mayer, Mautone, & Prothero (2002)	Profile Game (c)	Performance on retention and transfer tests	2	Content understanding, problem solving, and self regulation
Moreno & Mayer (2000)	Design-a-Plant (b)	Retention and transfer tests, plus survey	1, 2	Content understanding, problem solving, and self regulation
Parchman, Ellis, Christinaz, & Vogel (2000)	Game (a)	Retention tests	2	Content understanding
Porter, Bird, & Wunder (1990-1991)	Whale Game (a)	Performance on game, plus survey	1, 2	Collaboration, problem solving, and self regulation
Prislin, Jordan, Worchel, Semmer, & Shebilske (1996)	Space Fortress (a)	Performance on game, and observation	2, 4	Content understanding, collaboration, problem solving, communication, and self regulation

Table 2 (continued)

Study	Media[a]	Measures	Kirkpatrick levels[b]	CRESST categories
Rhodenizer, Bowers, & Bergondy (1998)	AIRTANDEM (b)	Performance on retention tests	2	Problem solving and self regulation
Ricci, Salas, & Cannon-Bowers (1996)	QuizShell (b)	Performance on pre-, post-, and retention tests	1, 2	Content understanding & self regulation
Rosenorn & Kofoed (1998)	Experiment Arium (b)	Observation	1, 2	Content understanding, collaboration, problem solving, communication, and self regulation
Shebilske, Regian, Arthur, & Jordan (1992)	Space Fortress (a)	Performance on game	1, 2, 4	Collaboration and problem solving
Tkacz (1998)	Map game (c)	Performance on transfer test	2	Content understanding, problem solving, and self regulation

[a] Letters in parentheses indicate type of media: a = game; b= simulation; c= simulation game.
[b] 1 = reaction at training event, 2 = learning at training event, 3 = behavior change on the job, and 4 = institutional benefits.

For the nine studies that used multiple measurements, six used two assessment methods, including at least one robust assessment method: game performance, a retention or transfer test, a knowledge map, or time to complete task. Another study utilized three assessment methods: retention tests, transfer tests, and a survey instrument.

Summary

In terms of Kirkpatrick's (1994) four levels of evaluation, in Table 2, six of the studies reviewed involved both Level 1 (reaction to training) and Level 2 (learning during training), nine studies involved Level 2 only (learning), only one study involved Level 3 (on-the-job changes due to training), and three studies involved Level 4 (benefits to the employer, for example: cost effectiveness). This suggests that while the Kirkpatrick model is appropriate for evaluating training programs, it is also a useful model for evaluating games studies. These results are different from those of the American Society for Training and Development (Sugrue & Kim,

2004) regarding the percentage of organizations using evaluation methods, in which the vast majority of evaluation methods were Level 1 (74%) and very few were Levels 2 (31%), 3, (14%) or 4 (8%). The difference between our study and the general training literature is most likely due to journal bias for significant results and our use of the criterion that there must be either qualitative or quantitative data.

The studies were also evaluated against the CRESST model of learning (Baker & Mayer, 1999). In terms of the CRESST model, 11 of the studies involved content understanding, 4 involved collaboration, 14 involved problem solving (domain specific, domain independent, or both), with 4 also involving communication or self-regulation (metacognition, motivation, or both). This suggests that while the CRESST model is appropriate for evaluating K-16 learning, it is also a useful model for evaluating games studies. These results further indicate that constructs investigated by empirical studies on the use of games to train adults are similar to the families of constructs that define the CRESST model of learning.

Because one of the major claimed advantages of games is that they are motivational, the existing CRESST framework needs to be augmented with an affective or motivational view of learning to be more useful for evaluating games and simulations. The current CRESST framework deals with motivation only in that the self-regulation component is composed of motivation (effort, self-efficacy), and the team skills component includes interpersonal skills as well as leadership, decision making, communication, adaptability. However, with these exceptions, the focus is clearly cognitive. Recent research offers some suggestions of what motivational factors to include. We have focused on those that predict academic achievement. For example, Robbins et al. (2004) provided an interesting basis for such a framework in a meta-analysis of the skill factors that predict college outcomes. They characterized motives as drives

(achievement motivation), goals (academic performance or mastery goals), and expectancies (self-efficacy and outcome expectations). We have added the motivational constructs of effort and anxiety, which are shown in Figure 3. In addition, based on Robbins et al. and our own research, we have provided definitions of these motivational constructs. These are shown in Table 3 (adapted from Robbins et al., p. 267). Other sources of research evidence for these motivational constructs can be found in Hidi and Harackiewicz (2000), Marsh, Hau, Artelt, and Baumert (2004), Pekrun, Goetz, Titz, and Perry (2002), and Seifert (2004).

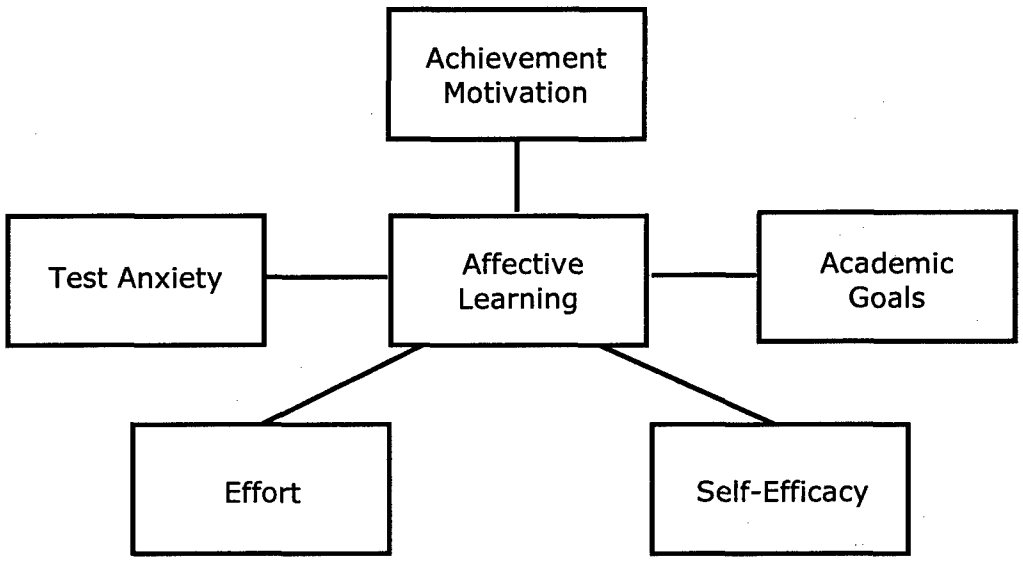

Figure 3. Affective/motivation model of learning.

Table 3

Motivation Constructs and Their Definitions

Construct	Definition
Achievement motivation	One's motivation to achieve success; enjoyment of surmounting obstacles and completing tasks undertaken; the drive to strive for success and excellence.
Academic goals	One's persistence with and commitment to action, including general and specific goal-directed behavior.
Academic self-efficacy	Self-evaluation of one's ability and/or changes for success in the academic environment.
Effort	O'Neil and Herl (1998) defined effort as the "extent to which one works hard on a task" (p. 1).
Test anxiety	The test anxiety literature (Hembree, 1988) categorizes test anxiety as both worry and emotionality. Worry is the cognitive concern about performance (e.g., "I wished I studied harder for this test"). Emotionality (or somatic anxiety) is a reaction to one's physiological responding (e.g., I feel my heart beating faster").

Source: Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130,* 261–288, p. 267.

Games, Simulations, and Assessments: A Convergence of Concepts

The field of educational measurement has enjoyed a long and productive history of developing practical methods for assessing what students know and can do. Concepts such as validity and comparability, methods such as item response theory and latent class analysis, and procedures such as theory-based task design and adaptive testing are the foundation of tests given to thousands of students each day, in myriad locations for widely varied purposes. However, recent developments in cognitive psychology and information technology reveal serious constraints on familiar testing procedures. They focus on individual students, minimize interactions with environments and other persons, and often deal with static situations.

There is a growing recognition of the central role in learning of ongoing interactions with environments and others (from both the information-processing and sociocultural stances within cognitive psychology). It has become possible to create such environments for students to interact in, as illustrated both by the emergence of coached practice environments such as

SHERLOCK (Katz & Lesgold, 1991) and multi-player real-time games such as Lunatix Online (Prowler Productions, www.prowler-pro.com). A question of vital interest to educational measurement is whether it is possible to extend the combination of rigor and utility in current assessments to more realistic assessments of a type better represented by a game than by a multiple-choice test item.

The angle of attack is a new line of research in educational assessment specifically intended to make explicit the underlying principles of educational assessment, and to develop an explicit design framework that can be applied to new as well as familiar kinds of assessments. This line of research is called "evidence-centered assessment design" (ECD; Mislevy, Steinberg, & Almond, 2003). This approach spans assessment arguments, an object-based design framework, and a delivery architecture consistent with shareable objects and processes compatible with international standards movements IMS/QTI and SCORM. Implementations of ECD object models and supporting design software are currently being developed under the NSF-supported project PADI—Principled Assessment Design for Inquiry, a collaboration among SRI International, the University of Maryland, UC Berkeley, and the University of Michigan < http://padi.sri.com/>

Groundwork for the proposed extensions has been laid in Mislevy's research program on the structure of assessment arguments (e.g., the ECD paper cited above), including an exploration of the extension of assessment arguments into interactive and group-based assessment scenarios in Mislevy (in press). Using Toulmin's (1958) structure for argumentation, we may display the structure of a familiar assessment task shown as Figure 4. Mislevy, Wilson, Ercikan, and Chudowsky (2003) reinterpret the psychometric principles of validity, reliability, comparability, and fairness as properties of assessment arguments structured in this manner.

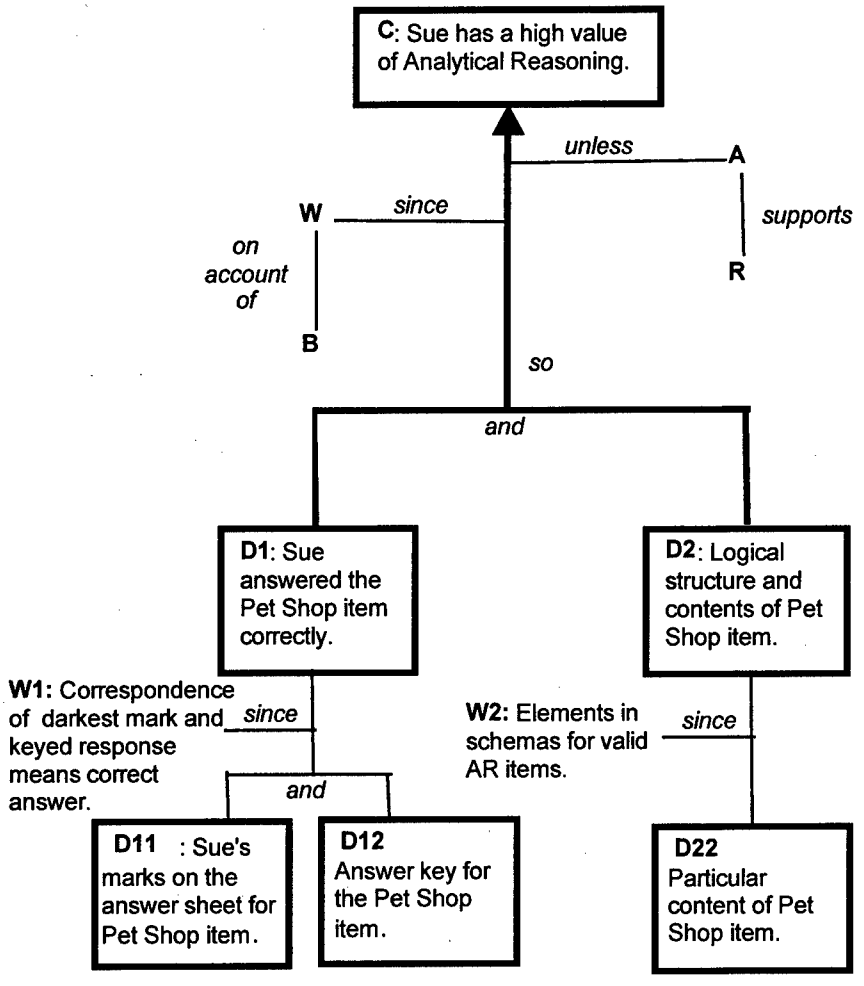

Figure 4. Elaborated Toulmin diagram for Pet Shop item. Adds detail to the process of reasoning from Sue's performance to the correctness of her answer and from the particulars of the Pet Shop item to its capability to evoke evidence about Analytical Reasoning ability. Note that a proposition such as D1 (Sue answered correctly) can be both a claim that depends on preceding propositions and (provisionally) an element of data for a subsequent claim.

Figures 5 and 6 show how this same approach to structuring an assessment argument extends to more complex tasks such as interactive troubleshooting in a simulation environment, and real-time conversation between two humans.

These argument structures would then be mapped into the "evidence-centered assessment design" (ECD) design layer in a framework such as the one being developed in the Principled Assessment Design for Inquiry (PADI) project. The project proposed here would be an analogous effort, developed and tested in collaboration with a team of experts in game environments.

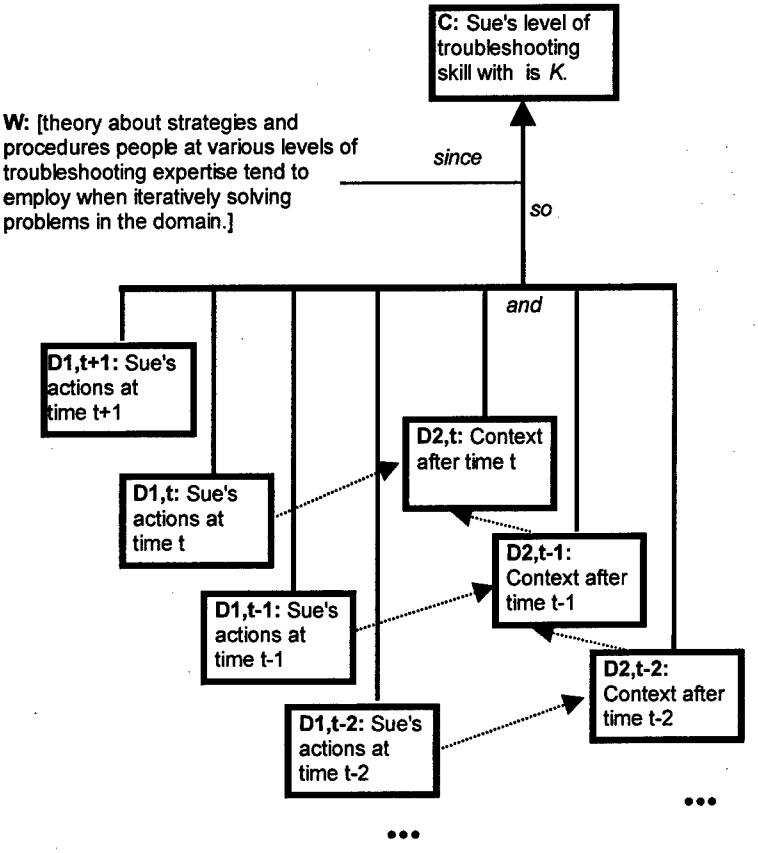

Figure 5. Elaborated Toulmin diagram for assessing hydraulics-system troubleshooting. Direct evidence for a claim about troubleshooting competence is obtained with a sequence of interactions between the examinee and the system, as the examinee proceeds through situation-hypothesis-action cycles. At each time point, the performance situation changes as a result of the examinee's previous actions and their effects on the system. Therefore the data from actions at different time points are serially independent.

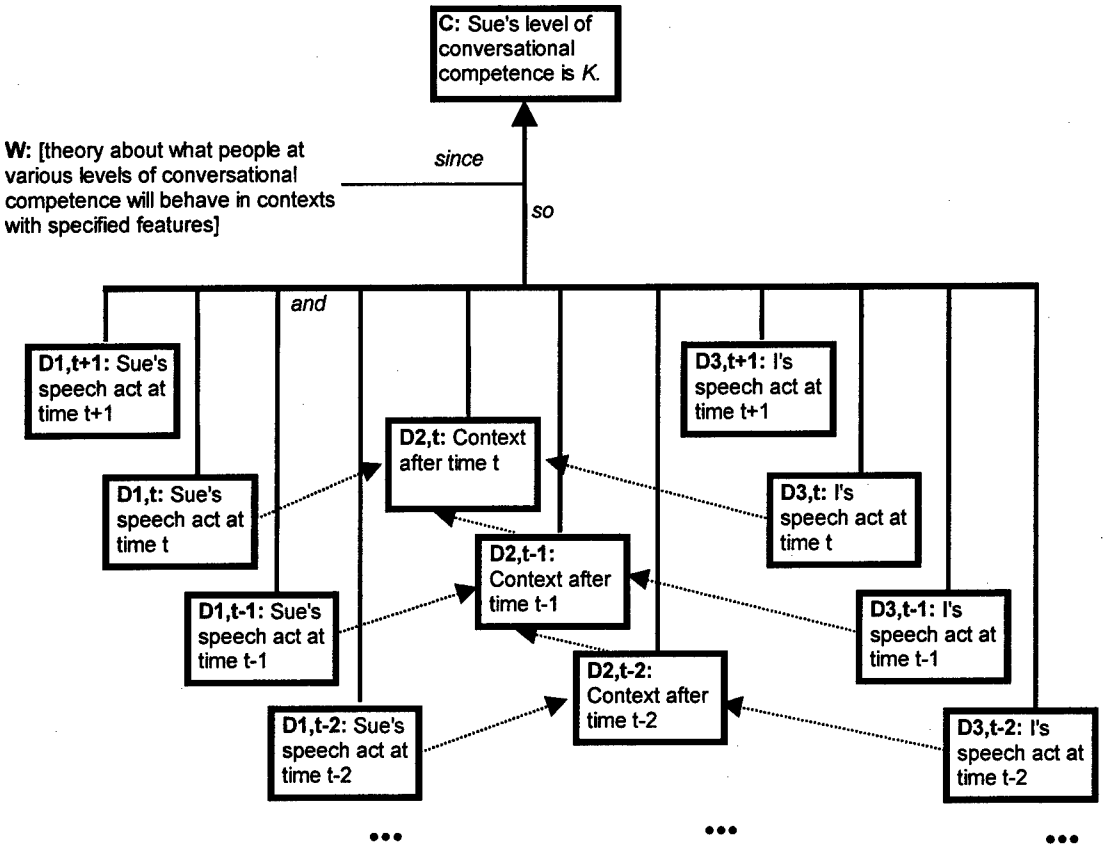

Figure 6. Elaborated Toulmin diagram for assessing conversational competence. Direct evidence for a claim about conversational competence is obtained through interactions between two or more people—two are addressed here. At each time point, the utterances of a person become part of a common performance situation, the context within which the next action must be evaluated. This figure concerns an oral interview, in which only a claim about the student's competence is desired.

Measurement Challenges

Game developers have low expertise in measurement, but high levels of game knowledge. Measures will need to be created that will be used in common across DARWARS projects. Examples include expert judgment on performance, training time to criterion, trainee background variables (e.g., prior knowledge, gender, experience in to-be-trained context), cost, and self-report

process measures indicating use and motivation. In addition, measures need to be created that will be unique to any given project (e.g., language facility). Common categories (but somewhat different measures) such as cost of failure, problem-solving components attempted and mastered, number of contexts experienced, etc. should be designed by DARWARS project staff.

Unique measures should be developed by the projects but should be reviewed by an assessment team. Is comprehensive and accurate information available on the trainee's background (e.g., quantitative and verbal aptitude, degree of prior knowledge, and experience in the training content)? Information on performance on the training task outcomes (e.g., quality of solution) as well as ongoing measures of behavioral (e.g., trainee's clickstream), conative (e.g., motivation) and cognitive and affective processes embedded within the task (e.g., measures of trainee understanding, stress) should be collected. A set of ideas that can lead to measurement opportunities (see Table 4) was provided by Dr. Patrick Kyllonen, one of our project consultants. As may be seen in Table 4, there is a range of measurement opportunities, from job-specific practice to management/administration.

Table 4

Measurement Opportunities

	Definition	Examples
Job-Specific Task Proficiency	Core tasks central to one's job	Knowledge of tactics
Non-Job-Specific Task Proficiency	Required tasks that cut across many jobs	Knowledge of chain of command; skill in seeking out information; skill in organizing a training session in DARWARS
Written & Oral Communication Task Proficiency	Formal oral and written presentations, independent of content	Writing a report summarizing a mission or sortie; presenting actions and results at an after-action review
Demonstration of Effort	Spending extra effort when asked, working in adverse conditions	Working on DARWARS late into the night and on weekends; logging hours on DARWARS
Maintenance of Personal Discipline	The degree to which negative behaviors are avoided	Avoiding alcohol and drug abuse, ethical infractions, excessive absenteeism
Facilitation of Peer and Team Performance	The degree to which one helps peers	Help others with advice, suggestions, in training simulations; involve others in group
Supervision/Leadership	Supervisory tasks involving face-to-face interaction and influence	Head a project, lead a DARWARS session; set goals for your team; reward and punish team members
Management/Administration	Supervision without face-to-face interaction	Set longer term goals for team; establish timelines, monitor progress; recruit resources

Evaluation Framework

The goals of our planning effort will be to operationalize specific evaluation questions, identify existing measures and metrics that could be employed, communicate with the providers of DARWARS software to assure the evaluation will be addressing the key intended outcomes, identify unintended outcomes, and propose a way to integrate findings across different implementations.

Conducting an evaluation of instructional technology has special challenges, particularly when different developers or designers are employed to create separate or interacting components

or courses. A critical step in an evaluation is to create a specification for game evaluation (see Table 5). As may be seen in Table 5, a series of decisions need to be made, from identifying specific hardware/software requirements to determining potential negative side effects.

Table 5

Specifications for Game Evaluation

General domain specification	Specific example
Specific hardware/software requirements	R&D system
Design issues	Evaluation vs. *research*
Content	
Scenario	Role playing
Participants	Novices, intermediates, experts
Instructional goals	
Type of learning	Collaboration, communication, *content knowledge, problem-solving*, self-regulation
Cognitive process (domain dependent)	*Prior knowledge*
Cognitive process (domain independent)	Visualization, dynamic spatial reasoning, situational awareness
Assessment strategies and measures	Multiple-choice tests, simulations, knowledge maps, essays
Reliability and validity issues	
Game issues	Fun, challenge, presence, immersion, fantasy
Affective processes (domain independent)	*Effort, worry, self-efficacy*
Instructional variables	Feedback (right/wrong, *visual, auditory*, explanatory, concurrent, immediate); analogies (visual, structural); *guided discovery* vs. expository; guided learning-by-doing; learner control with advisor; degree of scaffolding
Alignment of project assessment/evaluation with Architecture level of assessment/ evaluation	
Knowledge representation	Mental models
Formative evaluation	
Negative side effects	Impact on health/family

One set of evaluation questions focuses on whether the technology was used well or appropriately. A second set concerns whether the technology under development is supposed to

stabilize or is intended to be constantly adapted and upgraded as new software options occur. The latter choice obviously makes "summative" evaluation difficult, as the program is under constant redesign and development. Technology-based instruction as a target of evaluation also requires an understanding of the user(s), their technology experience, and the context of use. Simple problems like user verification vie in importance with complex issues such as knowledge retention in technology-based systems.

In specifying the evaluation of DARWARS, we recommend using the first two levels of Kirkpatrick's (1994) four-level evaluation model for formative evaluation. Initially, after success at Levels 1 and 2, then investigate Levels 3 and 4. The model focuses our effort on appropriate constructs of measurement. The four levels consist of the following. Level 1—Reaction: Measure of customer satisfaction. Do users react favorably to use of the system? Negative reactions reduce possibility of learning. Level 2—Learning: Extent to which participants change attitudes, improve knowledge, or increase skill. Level 3— Behavior: Extent to which change in behavior has occurred due to use of the system (i.e., transfer to the real world). Level 4—Results: Final results that occurred as participants used the system. Results can include improved quality, decreased costs, reduced accidents, increased sales, higher profits, and return on investment (ROI). In addition, the cost effectiveness of the system would be investigated. Measures should be developed based on a set of key questions that are asked before and after the use of the system. Example questions for Levels 1 and 2 are shown in Table 6.

We will suggest a formative evaluation approach to the evaluation (see Table 7). As may be seen in Table 7, a series of decisions need to be made, from checking the system design against its specifications to monitoring the implementation of revisions.

Table 6

Key Evaluation Questions for Users and Experts

1. Some Evaluating Reactions questions:
   - How do you rate the game?
   - How do you rate the training for the game?
   - How would you improve it?
   - Did it meet your needs?
2. Some Evaluating Learning questions:
   - What knowledge was learned?
   - What skills were developed or improved?
   - What standard of performance was set?
   - What attitudes were changed?

Note. Modified from: Kirkpatrick, D. L. (1994). *Evaluating training programs. The four levels.* San Francisco, CA: Berrett-Koehler.

Table 7

Formative Evaluation Activity

Check the system design against its specifications.
Check the design of assessments for outcome and diagnostic measurement against specifications. Design and try out measures.
Check the validity of instructional strategies embedded in the system using the *What Works* guidelines.
Conduct feasibility review with the instructors. – Are the right tasks being trained?
Conduct feasibility tests with the students, with groups of students, and the workforce. – One-on-one testing with protocol analysis – Small-group testing
Assess instructional effectiveness. – Cognitive, e.g., does it improve domain knowledge (for example, information technology skills, problem-solving skills, self-regulation)? – Affective, e.g., does it improve self-efficacy?
Do experts and novices differ in performance?
Does more training lead to better performance?
Monitor implementation of revisions.

Our Technical Approach to Planning

To accomplish our plan, we assembled a team of first-rate personnel to assist in developing the evaluation plan. The team consists of Harry O'Neil, University of Southern California, with experience in training, metacognition, and teamwork assessment, as well as planning; Robert Mislevy, University of Maryland, psychometrician with experience in high-technology assessment design and evaluation; Eva Baker, UCLA, evaluation and measurement expertise; Patrick Kyllonen, Educational Testing Service, whose work focuses on assessment of human abilities; and local Los Angeles talent from the game design environment (e.g., Richard Wainess). One of our activities was a needs assessment.

Needs Assessment

One of our first steps was to conduct an informal needs assessment at a 2003 DARWARS Conclave. Since one of the evaluation standards the Director of DARPA specified for the success of DARWARS was the willingness of the Services to cost-share, a question on this topic was asked at a DARWARS Conclave. The results of the needs assessment were surprising. We asked for a show of hands by government-only folks for "what evidence is needed to make a decision to cost-share a project." We asked for only "must have" evidence due to time constraints. Dr. Mike Freeman and the senior author counted hands on the following issues:

1. Trainee satisfaction (Mike's $n = 4$; Harry's $n = 5$)
2. Trainee learning (Mike's $n = 12$; Harry's $n = 15$)
3. Transfer of learning to the job (Mike's $n = 12$; Harry's $n = 13$)
4. Cost-effectiveness (Mike's $n = 9$; Harry's $n = 11$)

The numbers differ between the counters because they had different views of the meeting room.

What was surprising was that DARWARS (e.g., a game), which should be fun and thus should have high trainee satisfaction, was not very important to the decision-making process to co-fund. Not surprising was that transfer and cost-effectiveness were important. However, in DARWARS, neither will be measured. Another activity was the collection of information about DARWARS.

Information Collection

In order to collect information, site visits were conducted by Harry O'Neil at Information Sciences Institute (ISI), Acuitus, BBN Technologies, and MAK. Documentation (in the original form as emails to the DARPA brain trust) of these visits can be found in the earlier report. We also participated in DARWARS conferences to discuss our work and seek information. Finally, we visited the DARWARS booth at the Interservice/Industry Training, Simulation and Education Conference on December 9, 2004, to update our knowledge of DARWARS programs in terms of prototypes/demos available.

The Roadmap for Planning

This design document is provided in support of the evaluation of DARWARS program supported by DARPA. It contains sections that relate to the following tasks.

- Strategies about how to conduct the evaluation emphasizing cooperation and feasibility
- Technical design of the study
- Details of data collection
- Optional inclusion of comparative research designs as part of the formative evaluation
- Measurement framework, including existing, general, and specific measures and their analysis
- Draft of common reporting options.

Strategies about how to conduct the evaluation. The goals of the evaluation are to produce useful information for a number of constituents: program manager and advisory team; developers, users of the system, and the community at large. Each of these constituencies has different purposes for information, including time lines of implementation, comparisons, inferences to be drawn about the success of the system, and ways information collection itself can contribute to an effective learning environment. As the designers of the evaluation, our goals are to create a coherent model that can be used by other contractors or developers in either third-party or local formative evaluation of training products. All participants are interested in understanding the quality of their efforts and the means by which they might improve their programs.

Because evaluation is rarely seen in such a benign light, it is critical that our work be seen as complementing and supporting the work of the designers and developers of programs. Unless this view is held, we will be unlikely to get real cooperation in data collection, in overcoming inevitable barriers in empirical work, and in keeping to a plan that provides needed information on time and in a form that can be trusted. Two major sections are of interest: (a) the evaluation of LMTS applications; (b) the evaluation of the design and implementation of the multiuser game component.

The overall model for achieving the evaluation of LMTS requires an evaluation framework that makes both requirements and local options explicit. Thus, the evaluation involves components that are mandatory and optional. Within these components, there are mandatory and option measures. Let us begin with design concerns.

Mandatory design requirements. Two sets of evidence will be required to be provided. One will be longitudinal data by user across various tasks in the system. The second will be a comparative study of success in the system contrasted with other convincing comparisons.

Longitudinal data. The goal of this component is to provide evidence of learning patterns both in the process and interim acquisition of sub-objectives of the application. The LMTS provider must allow opportunities to collect data, potentially unobtrusively, about status and progress over time, in a manner that permits unequivocal tracing of individuals' trajectories. The data collaboratively identified should be classified as data relevant to the acquisition of the skillset or data related to the social, engagement, and affective states of the learner. While we recognize the interaction of these components, and recommend conducting analyses to explore how they are affected by characteristics of the training and individual differences, we want to avoid the dump of all data and the expectation that the evaluators alone will conduct signal-to-noise analyses. Should sufficient cases be available, growth modeling analyses, or simulations could be conducted.

Comparative data. Comparisons may be made cross-sectionally, at fixed points in time, or using longitudinal measures, if feasible. Ideally, trainees would be randomly assigned to treatment variations that include LMTS, comparative options, or no training, and pre and post measures would be obtained if they were not deemed to interact with subsequent training, or if they made sense in the particular environment. Random assignment may not be an option, in which case closely matched comparisons will be conducted with intact groups, where both the individual differences of the trainee and the learning strategies used in the comparison group would be managed, in order to draw explanations for data patterns. Criterion measures will involve those listed below.

Research design component. R&D providers should be encouraged to conduct comparative studies in the process of the development of their systems. These studies could include variables such as instructional approach, type and frequency of feedback, engagement in discretionary "enrichment" activities, and social and instructional supports outside of the computer environment. Learner control is not recommended. Providers with desire to conduct such studies will be supported by the evaluation contractor to the extent that studies generate data useful for formative evaluation of the system and findings of publishable quality.

Mandatory measures across all sites:

1. Background information on users
    - Education and training experience
    - Experience in technological environment in work
    - Experience in computer-supported training and distance learning
    - Background characteristics (e.g., ASVAB scores)
    - Self-report view of enterprise at initial contact

    For the purpose of the reporting, privacy of users will be protected, but temporary identifiers will be needed to collect data on process and outcomes.

2. LMTS domain-specific performance measures

    Each LMTS provider must provide the choice of process and outcome measures directly related to domain competency the provider is intending to develop and the schedule for administration of such measures. Using an exemplary form, the LMTS provider will answer questions about the degree to which the measures have existing validity evidence, and will also respond in detail about how performance is judged (i.e., criteria) and how performance is valued (i.e., cut scores or standards to be achieved).

These measures will be reviewed, where necessary using appropriate SMEs and feedback and advice will be provided to the provider within 30 days of receiving information.

3. Process measures

Also of interest is the design and collection of process measures intended to address the learning trajectories of individuals or teams. In order to accomplish this component, an inventory of process measures that LMTS or game architecture has designed will be accumulated. In addition, hypotheses about the meaning click stream patterns will be generated.

4. Domain independent analysis of LMTS measures.

To the extent possible, the cognitive requirements of the training tasks will be analyzed to see the degree to which they contain common forms of cognitive demands. To the extent this is found to be true following the application of a reliable rating system, then analysis techniques involving IRT approaches will be used to assess comparative progress across LMTS. A side effect of this approach will be a document that can help designers build into their architecture concern for desirable cognitive demands.

5. Self-report measures of process of competence acquisition will be employed.

Interviews with selected participants will be designed, and protocols created to probe issues related to perseverance, flaws or points of excellence in system design, needed support from other sources, motivation, barriers and so on.

Illustration by an application area: Rapid Tactical Language Training System. Recognizing that there will never be a highly competent linguist to support each unit or warrior conducting military operations overseas, DARWARS has initiated a project that is focused on spreading a limited, tailored knowledge of foreign language, culture, and gestures to *everyone*.

The goal is to deliver to every soldier or Marine a useful grasp of this "minimal" language in less than two weeks of training time, and to fill in the gaps where computer translators are weak: gestures, body language, and emotion. The program takes a new approach to second language instruction, delivering only those communication skills, competencies and background that are necessary to complete real-world missions and tasks. The Rapid Tactical Language Training System (RTLTS) places learners in a virtual world where they interact with animated characters representing local people and must perform face-to-face communication tasks in order to succeed in missions. They learn much more rapidly than with classroom instruction, because they get extensive practice employing their skills, as well as continual customized feedback aimed at their particular deficiencies. The application automatically collects usage data, which the RTLTS developers will use iteratively to improve the training system.

Evaluation requirements for the RTLTS:

1. Identify prerequisites for trainees.
2. Identify technology requirements and time frame.
3. Identify comparison groups, i.e., competent Arabic speakers, or first- and second-year cadets at West Point.
4. Identify independent segments of instruction.
5. Design longitudinal study, employing at least 30 trainees.
6. Implement criterion measures used by the military and language acquisition experts to examine validity data for certification and improvement purposes.
7. Design second criterion measure of language facility criterion measures, with individual oral testing.
8. Randomly assign trainees to training sequence, control, or alliterative practice; i.e., an individual tutor.
9. Implement affective measures in domain, with questions about sense of developing facility, comfort with cultural constraints.

10. Include the evaluation of physiological signs to predict learning.
11. If data are robust, conduct multivariate analyses addressing comparative strengths, patterns of acquisition, and inferences by developers from data.

In general the intention of the analyses would be to provide accurate status estimates of level of competency attained by trainees, moderated where possible by individual background variables, observed motivation, process variables, and level of engagement in learning. Both quantitative and qualitative data would be analyzed. Hierarchical models would be used where there are sufficient numbers, in particular to estimate individual and team effects of training.

In addition, techniques to determine the relative cross application effects would be estimated using Item Response Theory, and Bayesnets. The overall intent of the data analysis would be to characterize the learning process and states attained by trainees over time, to compare their status with other scientifically selected and assigned comparison groups, to determine which measures are most sensitive to implementations, and to provide a synthesis of the findings of the evaluation in terms of trustworthy results and validity of the inferences made by developers.

Reports would be presented interactively and graphically to illustrate simultaneously the degree of progress and attainment, Each report would have common features located in identical places on the page. Common icons would be used across implementations to indicate learners, instructional time, dependent measures, and value of findings. Drill down opportunity would be possible by clicking on performance to learn more about measures, program, or analytical techniques. Summary comments about results would be made and recommendations would conclude the display.

The final component of such a plan would list the cost for implementing the evaluation. The schedule for conduct of the evaluation would also be detailed. The plan would provide a blueprint for the design and aggregation of reports of the effectiveness of the system, by tasks,

users, level of use, and degree of difficulty. Report templates would be prepared to illustrate audience-adapted reports.

Availability of Editors

In order to embed assessment measures, collect process data, and provide an opportunity for changing installation strategies, some means of changing the off-the-shelf game must be available. Editors that are available for some games are one mechanism. Thus, we investigated the availability of level editors in the following manner. The off-the-shelf game needed to be (a) Windows XP compatible, (b) able to be learned in 1/2 hour or less, (c) using current or modern weapons, (d) real-world, and (e) on par with current video game quality. It was also helpful if it handled both single and multiplayer modes.

There were three phases in the search for games with suitable editors, with Phase 1 containing two parts. Phase 1–Part 1 involved approximately 90 pages of data tables. Phase 1–Part 2 involved approximately 70 pages of data tables. Phase 2 involved 9 pages of data tables. Phase 3 includes 3 pages of data tables. The results of the analysis for Phases 2 and 3 can be found in Tables 8 and 9. The results of Phase 1 will be provided by the authors upon request.

Phase 1–Part 1 resulted in selection of 91 games with humans as enemies. Phase 1–Part 2 resulted in selection of 73 of those games by removing any containing fantasy, myth, or ancient themes. Phase 2 resulted in selection of 9 of the games by retaining only those involving modern warfare and playable on Windows XP. Phase 3 resulted in selection the top 3 games, based on (a) retaining only those that could be learned in 1/2 hour or less and (b) of those, the ones with the highest ratings from multiple sources. The final game chosen was Soldiers of Anarchy.

Table 8

War Games With Level Editors—2nd Phase: Selection of 9 War Games

WAR GAMES WITH LEVEL EDITORS
2ND PHASE: SELECTION OF 9 WAR GAMES WITH HUMANS AS ENEMIES
(retained only games involving modern warfare and could be played on Windows XP)

www.epinions.com Searched Tuesday June 2, 2004

Computer games (7100 listed)
Strategy and War Games (520 listed)
Selected games with humans as enemies
Retained only games with level editors, involved modern warfare, and could be played on XP

TITLE	Yr. Rel.	Rating (max 5)	Time to learn	Demo	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail
Hearts of Iron for Windows	Nov, 2002	3.9	2.5 hours		Yes	RTS	Top	X	X	yes	98, me, 2k, NT, XP	yes

Review URL	Time frame	Description
http://www.gamershell.com/reviews/HeartsofIronReview.shtml http://hardwarecentral.dealtime.com/xPF-Hearts_Of_Iron http://www.mobygames.com/game/sheet/gameId=7953 http://www.netjak.com/Reviews/windows/hoi.htm http://www.gamespot.com/pc/strategy/heartsofiron/review.html	WWII	Hearts of Iron is one of the first WWII PC games to encompass such a grand strategic scope. The game map spans the entire world, including all nations between 1936 – 1946. The game allows players to take the War to any new front of their own choosing, while focusing primarily on the epic struggle between the great alliances of the New World Orders – Fascism, Communism and Democracy. Players are able to modify and construct their own settings and campaigns with the game's scenario generator, which offers thousands of combinations, not to mention hours of gameplay. Developed by the same team behind the award winning Europa Universalis series, Hearts of Iron features a highly advanced research model that allows players to acquire more powerful weapons through the course of the war; a unique political system that includes historical generals and political leaders that has a direct effect on neutral nations; over 100 different ground, air and naval forces; and multiplayer mode over the Internet.

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Fleet Command for Windows	May, 1999	3.5, 3.75	3 hours		Yes	Strategy/ Sim	Top/ 2nd	X	max 4		NT, 2K, XP	Yes

Review URL	Time frame	Description
http://hardwarecentral.dealtime.com/xPF-Fleet_Command http://hardwarecentral.dealtime.com/xPR-Fleet_Command~RD-1201430 http://www.gamewinners.com/DOSWIN/bljanesfleetcommand.htm http://www.3dgamers.com/games/fleetcommand/ http://www.sonalysts.com/entertainment/ent_fleet_command.html http://www.gamespot.com/pc/strategy/janesfleetcommand/	modern	Some games let you command a single airplane or one ship, but Jane's Fleet Command puts the power (and fate) of an entire naval task force in your hands. It's up to you to schedule air strikes, protect your vital aircraft carriers and reconnaissance aircraft, and orchestrate air coverage so enemy units never have a chance to penetrate your defenses. You'll give orders to submarines, destroyers, air bases, and every other major battle platform in the particular combat theater you're fighting through. Normally a game like this would play out from a standard overhead map, but Fleet Command immerses players in the battle by depicting everything in a 3-D environment. Hear that distress call from a lonely reconnaissance plane? Zoom in to watch the enemy chase it down and (hopefully) see your own aircraft scream to the rescue. It's possible to follow the movements of any object in the game, rotating the camera until you find the perfect angle. It's almost like being the director of your own war.

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Soldiers of Anarchy	Oct., 2002	3.5, 4.0, 3.75	half hour	Yes	Yes	RTS, squad-based	1st	X	max 8		98, me, 2k, XP	Yes

Review URL	Time frame	Description
http://www.soldiers-of-anarchy.com/ http://www.silver-style.com/e/games/e_soa.html http://www.3dgamers.com/games/soldiersanarchy/ http://www.gamezone.com/gamesell/p20936.htm http://www.actiontrip.com/reviews/soldiersofanarchy.phtml http://www.dailygame.net/Articles/Reviews/soldiers.html http://www.fileplanet.com/files/110000/112223.shtml http://www.game-over.net/reviews.php?id=776 http://www.gamespot.com/pc/strategy/soldiersofanarchy/ http://www.3dgamers.com/news/more/1030460833/ http://www.entdepot.com/pc/soldiersofanarchy/review.php	future, post apocalyptic	Fight against the worst enemy mankind has ever faced...itself! It's 2013... 10 years ago, a fanatical cult caused a near apocalypse on Earth. To survive, you and a group of military veterans went underground and now you have emerged to a world enslaved.

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Combat Mission 2: Barbarossa to Berlin for Mac, Windows	Sept., 2002	4.7		Yes	Yes	Turn-Based, Strategy, RT	3rd	X	yes		95, 98, me, 2k, XP	
Review URL						**Time frame**	**Description**					
http://archive.gamespy.com/reviews/november02/combatmissionbbpc/ http://www.software-reviews.net/Barbarossa/review/software/0688042686 http://games.tucows.tierra.net/preview/282686.html http://www.battlefront.com/products/cmbb/demo.html http://www.battlefront.com/products/cmbb/readme.html http://archive.gamespy.com/games/5185.shtml http://www.worthplaying.com/article.php?sid=7656						WWII	COMBAT MISSION 2: BARBAROSSA TO BERLIN ventures into the heart of World War II wargaming: the Eastern Front. The struggle between Nazi Germany and the Soviet Union captures the flavor of the four years of conflict that left Germany in ruins. Combat Mission 2 includes over 125 battles between 1941 and the final capitulation in May 1945, and many more are possible with the full featured editor. You are the commander of breathtaking battles set in Moscow, Stalingrad, Berlin and many more well-known battlefields. Combat Mission 2: Barbarossa to Berlin (Special Edition) involves a variety of units and tactical possibilities that know no competition—prepare for war on the Eastern front!					

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Year Released	Rating (max 5)	Time to learn	Demo	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Conflict Zone for Windows	Oct., 2001	2.8, 2.1	half hour	Yes	Yes	RTS, RPG	Iso	X	max 8		95, 98, me, 2k, XP	Yes

Review URL	Time frame	Description
http://vgstrategies.about.com/library/doswin/blconflictzone.htm http://www.3dgamers.com/games/conflictzone/ http://www.3dgamers.com/news/more/987943340/ http://www.worlddesign.net/pc_cheatss/Conflict_Zone.html http://www.gamewinners.com/DOSWIN/blconflictzone.htm http://www.mobygames.com/game/sheet/gameId%2C6198 http://www.3dgamers.com/games/conflictzone/	modern day	Modern-day warfare collides with public perception in Conflict Zone, a solid military strategy game that rewards the media-savvy commander. With TV cameras trained on your every move, public perception holds the key to your military future. Make your country proud and you could be promoted. Botch your assignment and you could lose your command. Gather your resources, build your bases, and fight your battles as either the international corps for peace or the power-hungry dictators. Plan carefully—the world is watching.

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
War Times for Windows	April, 2004	2.5, 3.8, 2.6, 2, 3.9	0 to 15 minutes	Yes	Yes	RTS	God	X	max 14		98, me, 2k, XP	Yes
Review URL						**Time frame**	**Description**					
http://www.3dgamers.com/games/wartimes/ http://www.strategyfirst.com/games/GameInfo.asp?sLanguageCode=EN&iGameID=103&sSection=Overview http://www.gamezone.com/gamesell/p22143.htm http://www.gamespot.com/pc/strategy/wartimes/preview_6075566.html http://www.gamezone.com/gamesell/hints/h22143.htm http://www.gamerankings.com/htmlpages2/914817.asp						WWII	War Times is a real-time strategy game from Strategy First and Legend Studio that attempts to capture the intricacies of the struggle for position and power during World War II. To that end, the game employs many of the elements of the typical RTS title, but also breaks out with some unique perspectives of the titanic struggle by employing a host of controllable units, coupled with a solid options package and graphics set.					

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Massive Assault for Windows	**Oct., 2003**	**3.6, 4.4**	2 hours	Yes	coming	turn-based 3D	God	X	yes		98, me, 2k, XP	no

Review URL	Time frame	Description
http://www.netjak.com/review.php/517 http://www.3dgamers.com/games/massiveassault/ http://www.gamescore.com/index.php?action=news&id=129 http://www.gamespot.com/pc/strategy/massiveassault/downloads.html http://www.gamespot.com/pc/strategy/massiveassault/news_6085972.html http://www.gamespot.com/pc/strategy/massiveassault/index.html http://www.gamescore.com/index.php?action=news&id=129	future	Can you save the future? Massive Assault utilizes Wargaming.net's high-performance 3D engine giving a revolutionary look and feel to turn-based, global-scale games. Gameplay is smooth and easy-to-learn.

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Call of Duty for Windows	Nov., 2003	4.8, 4.5	half hour	Yes	Yes	RTS, tactical	1st	X	yes		98, me, 2k, XP	Yes
Review URL						Time frame	Description					
http://hardwarecentral.dealtime.com/xPR-Call_Of_Duty~RD-121531436676 http://www.netjak.com/review.php/479 http://www.cheatpatch.com/index.php?action=search&ID=7622&gamename=Call+of+Duty http://www.gamewinners.com/DOSWIN/blcallduty.htm http://www.computer-games-station.com/video-games/9734.htm http://www.3dgamers.com/games/callofduty/ http://www.gamespot.com/pc/action/callofduty/hints.html?tag=gs_tabht http://www.youplayimod.com/Tutorials-list-4.html http://www.gamespot.com/pc/action/callofduty/index.html http://www.gamespot.com/pc/action/callofduty/downloads.html						WWII						

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
FarCry	Mar., 2004	4.5, 5.0, 4.6	0 to 15 minutes	Yes		RTS	1st	X	max 8		98, me, 2k, XP	
Review URL						Time frame	Description					
http://www.computer-games-station.com/video-games/9915.htm http://farcry-thegame.com/uk/system.php http://www.gamespot.com/pc/action/farcry/ http://www.gamezilla.com/review.aspx?review=8835 http://farcry.ubi.com/demo.php http://www.3dgamers.com/games/farcry/ http://www.pcreview.co.uk/shop-FarCry-item_id-B00009LW88-search_type-AsinSearch-locale-uk.php http://www.gamespot.com/pc/action/farcry/index.html												

Table 8 (continued)

2ND PHASE: SELECTION OF 9 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Full Spectrum Warrior	June, 2004? Sept., 2004	3.5, 3.3, 3.7				RTS, tactical, squad	1st	X	Yes			
Review URL						Time Frame	Description					
http://www.fullspectrumwarrior.com/ http://www.3dgamers.com/games/fullspecwarrior/ http://www.gengamers.com/html/full_spectrum_warrior.html http://www.ugo.com/channels/games/features/fullspectrumwarrior/ http://archive.gamespy.com/games/6453.shtml http://												

Table 9

3rd Phase: Selection of 3 War Games With Level Editors

3RD PHASE: SELECTION OF 3 WAR GAMES WITH LEVEL EDITORS
www.epinions.com Searched Tuesday June 2, 2004 Computer games (7100 listed) Strategy and War Games (520 listed) Selected games with humans as enemies In **third phase**, retained only games with level editors, involved modern warfare, could be played on XP, could be learned in a half hour or less, and received high ratings from game reviewers. Note: SP = Single Player, MP = Multiplayer, MMP = Massively Multiplayer, O/S = operating system

TITLE	Yr. Rel.	Rating (max 5)	Time to learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Soldiers of Anarchy	Oct., 2002	3.5, 4.0, 3.75	half hour	Yes	Yes	RTS, squad-based	1st	X	max 8		98, me, 2k, XP	Yes

Review URL	Time frame	Description
http://www.soldiers-of-anarchy.com/ http://www.silver-style.com/e/games/e_soa.html http://www.3dgamers.com/games/soldiersanarchy/ http://www.gamezone.com/gamesell/p20936.htm http://www.actiontrip.com/reviews/soldiersofanarchy.phtml http://www.dailygame.net/Articles/Reviews/soldiers.html http://www.fileplanet.com/files/110000/112223.shtml http://www.game-over.net/reviews.php?id=776 http://www.gamespot.com/pc/strategy/soldiersofanarchy/ http://www.3dgamers.com/news/more/1030460833/ http://www.entdepot.com/pc/soldiersofanarchy/review.php	future, post apocal-yptic	Fight against the worst enemy mankind has ever faced...itself! It's 2013... 10 years ago, a fanatical cult caused a near apocalypse on Earth. To survive, you and a group of military veterans went underground and now you have emerged to a world enslaved.

Table 9 (continued)

3RD PHASE: SELECTION OF 3 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr. Rel.	Rating (max 5)	Time to learn	Demo Avail.	Map Edit	GameType	Perspec-tive	SP	MP	MMP	O/S	Cheats Avail.
Combat Mission 2: Barbarossa to Berlin for Mac, Windows	Sept., 2002	4.7		Yes	Yes	Turn-Based, Strategy, RT	3rd	X	yes		95, 98, me, 2k, XP	
Review URL						**Time frame**	**Description**					
http://archive.gamespy.com/reviews/november02/combatmissionbbpc/ http://www.software-reviews.net/Barbarossa/review/software/0688042686 http://games.tucows.tierra.net/preview/282686.html http://www.battlefront.com/products/cmbb/demo.html http://www.battlefront.com/products/cmbb/readme.html http://archive.gamespy.com/games/5185.shtml http://www.worthplaying.com/article.php?sid=7656						WWII	COMBAT MISSION 2: BARBAROSSA TO BERLIN ventures into the heart of World War II wargaming: the Eastern Front. The struggle between Nazi Germany and the Soviet Union captures the flavor of the four years of conflict that left Germany in ruins. Combat Mission 2 includes over 125 battles between 1941 and the final capitulation in May 1945, and many more are possible with the full featured editor. You are the commander of breathtaking battles set in Moscow, Stalingrad, Berlin and many more well known battlefields. Combat Mission 2: Barbarossa to Berlin (Special Edition) involves a variety of units and tactical possibilities that know no competition – prepare for war on the Eastern front!					

Table 9 (continued)

3RD PHASE: SELECTION OF 3 WAR GAMES WITH LEVEL EDITORS (cont'd)												
TITLE	Yr Rel.	Rating (max 5)	Time To learn	Demo Avail.	Map Edit	Game Type	Per-spec-tive	SP	MP	MMP	O/S	Cheats Avail.
Call of Duty for Windows	Nov., 2003	4.8, 4.5	half hour	Yes	Yes	RTS, tactical	1st	X	yes		98, me, 2k, XP	Yes
Review URL						Time frame	Description					
http://hardwarecentral.dealtime.com/xPR-Call_Of_Duty~RD-121531436676 http://www.netjak.com/review.php/479 http://www.cheatpatch.com/index.php?action=search&ID=7622&gamename=Call+of+Duty http://www.gamewinners.com/DOSWIN/blcallduty.htm http://www.computer-games-station.com/video-games/9734.htm http://www.3dgamers.com/games/callofduty/ http://www.gamespot.com/pc/action/callofduty/hints.html?tag=gs_tabht http://www.youplayimod.com/Tutorials-list-4.html http://www.gamespot.com/pc/action/callofduty/index.html http://www.gamespot.com/pc/action/callofduty/downloads.html						WWII						

Program Prototype Assessment Into a Game

Although programming a prototype assessment into a game was not feasible with this project's resources (due to lack of funding), a dissertation chaired by Harold O'Neil and conducted by Claire Chen at the University of Southern California investigated the feasibility of the approach. This section of the report is the abstract of a presentation by Chen and O'Neil that has been accepted for presentation at the 2005 annual meeting of the American Educational Research Association.

Purpose and Theoretical Framework

Will adults increase their problem-solving skills after playing a computer game? This study evaluated a computer game (SafeCracker by Dreamcatcher Interactive, Inc.) with regard to its effectiveness in improving problem solving using the problem-solving assessment model developed by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) to measure content understanding, problem-solving strategies, and self-regulation, the three elements of problem-solving ability (Herl, O'Neil, Chung, & Schacter, 1999; Mayer, 2002). A pilot study focused on a formative evaluation to improve implementation and ensure the success of the main study. The main study focused on the impact of the game on problem-solving skills.

As defined by Gredler (1996), "games consist of rules that describe allowable player moves, game constraints and privileges (such as ways of earning extra turns), and penalties for illegal (nonpermissible) actions" (p. 523). In addition, the rules of games do not have to obey those in real life and can be imaginative.

As noted by Ruben (1999), researchers began to notice the potential effects of simulations and games in instruction decades ago and applied them in various fields and settings, such as

business and industry, K-16 organizations, and military organizations (e.g., Donchin, 1989; Malone, 1981; Quinn, 1991; Ruben, 1999; Thomas & Macredie, 1994), and in various subjects, such as geology (Mayer et al., 2002; Moreno & Mayer, 2000; Tkacz, 1998), business (e.g., King & Morrison, 1998; Schank, 1997), physics (e.g., White & Frederiksen, 1998), and in therapeutic situations (Ruben, 1999). However few studies using adults as participants have shown empirically the learning effects of games (O'Neil & Fisher, 2004).

According to previous research, problem solving is one of the most critical competencies, whether for lifetime learning or accomplishing tasks, whether in job settings, academic settings (e.g., Dugdale, 1998), or any other setting. Although there is substantial previous research that reveals the utility of problem solving (e.g., Mayer, 2002), current computer-based methods to assess problem-solving skills still need to be refined. As argued by Quinn (1991), computer games may provide effective environments for measuring problem solving.

One of the most important supporting theories for using games and simulations is found in experience-based learning (Ruben, 1999). Experience-based learning is an important approach focusing on increasing the student's control and autonomy, an aspect of constructivism. Experience-based learning is the process whereby students learn something by doing it and connecting it with their prior experience, such as in hands-on laboratory experiments, studio performances, practical training, etc. Computer games and simulations could facilitate experience-based learning by transporting learners to "another world," where they are in control of the action, and by providing them opportunities to interact with a knowledge domain (Gredler, 1996). However, there are limited experimental data (Clark, 2001; Mayer, 2004; Sweller, 1994) showing that such an approach facilitates learning in general or learning from games in particular.

Another characteristic of games is that they use an inductive/discovery learning strategy in which learners acquire knowledge by exploring in environments, such as simulations (Rieber & Parmley, 1995), by themselves with little or no guidance (Kalyuga, Chandler, Touvinen, & Sweller, 2001). Discovery learning is a common instructional method in games. However, experimental evidence (Clark, 2001; Mayer, 2004; Sweller, 1994) also indicates that an inductive/discovery learning strategy may not be an effective instructional strategy. Thus, our prediction is that without an explicit effective instructional strategy, off-the-shelf games may not facilitate learning. In the pilot study, our first task was to create instrumentation in a game context to measure problem solving. We then conducted the main study.

Pilot Study Methods and Techniques

For the pilot study, we applied a framework of formative evaluation. The pilot study followed a modified version of the methodology of O'Neil, Baker, and Fisher (2002) to conduct a formative evaluation of a game (Table 10).

Table 10

Formative Evaluation Activity (adapted from O'Neil et al., 2002)

Check the design of assessments for outcome and measurement. Design and try out measures.
Check the validity of instructional strategies embedded in the game against research literature.
Conduct feasibility review with the students.
Implement revisions.

The participants in the pilot study were three college students, ages 20 to 35 years, from a university in northern California. The pilot study was conducted after receiving approval of the research university's IRB. All participants were selected to have no experience in playing the computer game SafeCracker.

SafeCracker is a computer puzzle-solving game in which a player is an applicant for a job as the head of security development at a world famous firm. As part of the application process, the player needs to try to break into the safes located in a mansion, finding clues and using every available resource.

The selection of SafeCracker was based on results from a study by Wainess and O'Neil (2003), who evaluated the potential research feasibility of 525 video games. An appropriate game was then sought among puzzle games, due to their problem-solving properties and because they provide an appropriate platform for studying the effectiveness of games for enhancing problem-solving skills. A participant in a puzzle-solving game is placed in a specific setting or story background and tries to reason out possible task procedures and consequences. Failure to solve a puzzle previously encountered may result in future problems in the game. The study used three sets of problem-solving measures.

Content understanding measure. *Knowledge Mapper* is a computer-based concept mapping instrument that was used to measure participants' content understanding of SafeCracker. The Knowledge Mapper used in previous studies (e.g., Chuang, 2004; Hsieh, 2001; Schacter, Herl, Chung, Dennis, & O'Neil, 1999) was reprogrammed to fit the needs of this study. Participants in the study were asked to play SafeCracker twice and after each game to create a knowledge map in a computer-based environment. Participants were evaluated on their content understanding, based on their maps, after both the first and the second time of playing SafeCracker. Participants' maps were scored in real time by comparing the semantic propositions of a participant's knowledge map to those of three expert players' maps. For example, if a participant made a proposition such as "Key is used for safe," this proposition would be then compared with all of the propositions in the three expert maps.

Problem-solving strategies measure. In this study, the researchers measured domain-specific problem-solving strategies by asking open-ended questions, using modifications of previous researchers' (e.g., Mayer, Dow, & Mayer, 2003; Mayer & Moreno, 1998; Mayer, Sobko, & Mautone, 2003) assessments of retention and transfer. For example, the researchers in this study adapted Mayer and Moreno's (1998; also, Mayer, Sobko, & Mautone, 2003) approach to measure a participant's retention and transfer by counting the number of predefined major idea units correctly stated by the participant regardless of wording. Two problem-solving strategy questions to measure retention and transfer were used in the pilot study. The retention question was "Write an explanation of how you solve the puzzles in the rooms." The transfer question was "List some ways to improve the fun or challenge of the game." An example of an idea unit for the problem-solving retention question is "Recognize/compare room features." An example of an idea unit for the transfer question is "Add characters to disturb/confuse or help."

Self-regulation measure. The trait self-regulation questionnaire designed by O'Neil and Herl (1998) was used in this study to assess participants' degree of self-regulation, one of the components of problem-solving skill. The reliability of the self-regulation questionnaire ranged from .89 to .94 (O'Neil & Herl). The 32-item questionnaire had four factors—planning, self-checking, self-efficacy, and effort—with 8 items each. An example of an item to assess planning ability is "I determine how to solve a task before I begin." An example of an item to assess self-efficacy is "I check how well I am doing when I solve a task." Item response choices were *almost never* (1), *sometimes* (2), *often* (3), and *almost always* (4).

Procedure. Participants in the pilot study were tested individually on a PC. The process was: (a) 2-3 minutes for study introduction, (b) 8 minutes to complete the self-regulation questionnaire, (c) 8 minutes for instruction on knowledge mapping, (d) 5 minutes for game

introduction, (e) 20 minutes for the first game session, (f) 5 minutes for the first knowledge map, (g) 4 minutes for the first problem-solving strategy questions, (h) 20 minutes for the second game session, (i) 5 minutes for the second knowledge map, (j) 4 minutes for the second problem-solving strategy questions, (k) 2 minutes for debriefing. The entire session took a maximum of 90 minutes.

The researcher briefly introduced the SafeCracker game to the participants, including the gist and mechanics of the game. In addition, the participants were told the tasks they were going to perform and that they needed to try every way and use every available resource to show their capability by solving the puzzles and cracking the safes located in the game's rooms.

Main Study

Thirty young adults, ages 20 to 35 years, participated in the main study. All participants were selected to have previous experience in playing a computer puzzle-solving or strategy game, but no experience in playing SafeCracker. Participants were paid for participating. Except for minor adjustments, the same methods and procedure used in the pilot study were applied in the main study.

Based on the pilot study results for the problem-solving strategy test, the transfer question "Write an explanation of how you solve the puzzles in the rooms" was revised as "List how you solve the puzzles in the rooms" to reduce the probability that participants would write an essay. In this study, the goal was for participants to write down only key words and phrases, not only to save time but also to make the idea units/propositions clear.

Data sources and results. The results showed a significant increase, $t(29) = 4.32$, $p < .05$, in overall content understanding after playing SafeCracker ($M = 3.44$, $SD = 1.84$) compared with before playing the game ($M = 2.27$, $SD = 1.23$).

The results also showed that participants' problem-solving strategies significantly increased, $t(29) = 12.66$, $p < .05$. For example, participants' overall retention was significantly better ($M = 4.43$, $SD = 1.52$) after the second round of playing SafeCracker than after the first round ($M = 2.46$, $SD = 1.13$). Additionally, Cohen's Kappa coefficient (the percentage of agreement between the two raters in scoring for problem-solving retention) was .95.

The results also showed that participants performed significantly better on the test of transfer, $t(29) = 7.05$, $p < .05$; that is, participants' problem-solving strategy for transfer was significantly better ($M = 2.76$, $SD = 1.47$) after the second round of playing SafeCracker than after the first round ($M = 1.70$, $SD = 0.98$). In addition, Cohen's Kappa coefficient for the transfer measure was .95. For the self-regulation process, no significant relationship was found. This result was expected as the trait self-regulation questionnaire was administered before the game.

It should be noted that the overall level of game playing was low on the posttest; when compared with the experts' maps, students' maps included only 4% of expert knowledge. However, this result was not unexpected, as no effective instructional strategies were provided in the game.

Summary

Despite computer games and simulations' potential power in instruction and training, research on their training effectiveness for adults is limited. The researchers conducted a formative evaluation of a computer game in terms of its effectiveness for enhancing learners' problem-solving skills, including content understanding, domain-specific problem-solving strategies, and self-regulation. A pilot study and a main study were conducted. The pilot study was conducted to try out the measures and procedures. The results of the pilot showed that the study was feasible but required several revisions. The main study results showed that playing the

computer game enhanced participants' problem-solving skills significantly from pretest measures to posttest measures; however, their performance was low compared to that of expert game players.

In this study, we conducted a formative evaluation and assessment of whether students' problem-solving skills increased due to playing the SafeCracker game. Scientifically it is one of a few studies to explicitly measure game problem solving and the only one to use expert-based knowledge maps in this context. From an educational point of view, the study provides a caution that using games "as is" (off-the-shelf) without an explicit instructional design may not lead to educational benefits.

A Closing Comment: Relationship of Instructional Design to Effective Games

Our position is that games themselves are not sufficient for learning, but there are elements in games that can be activated within an instructional context that may enhance the learning process (Garris, Ahlers, & Driskell, 2002). In other words, outcomes are affected by the instructional strategies employed (Wolfe, 1997). Leemkuil, de Jong, de Hoog, and Christoph (2003), too, commented that there is general consensus that learning with interactive environments such as games, simulations, and adventures is not effective when no effective instructional measure or support is added.

de Jong and van Joolingen (1998), after reviewing a large number of studies on learning from simulations, concluded, "There is no clear and unequivocal outcome in favor of simulations. An explanation why simulation based learning does not improve learning results can be found in the intrinsic problems that learners may have with discovery learning" (p. 181). These problems are related to processes such as hypothesis generation, design of experiments, interpretation of data, and regulation of learning. After analyzing a large number of studies, de Jong and van

Joolingen concluded that adding instructional support (i.e., scaffolding) to simulations might help to improve the situation. A similar conclusion can be applied to games.

According to Thiagarajan (1998), if not embedded with sound instructional design, games and simulations often end up truncated exercises often mislabeled as simulations. Gredler (1996) further commented that poorly developed exercises are not effective in achieving the objectives for which simulations are most appropriate—that of developing students' problem-solving skills. According to Lee (1999), for instructional prescription, we need information dealing with instructional variables, such as instructional mode, instructional sequence, knowledge domain, and learner characteristics.

There appears to be consensus among a large number of researchers with regard to the negative, mixed, or null findings of games research, suggesting that the cause might be a lack of sound instructional design embedded in the games (de Jong & van Joolingen, 1998; Garris et al., 2002; Gredler, 1996; Lee, 1999; Leemkuil et al., 2003; O'Neil & Fisher, 2004; Thiagarajan, 1998; Wolfe, 1997).

An important component of research on the effectiveness of educational games and simulations is the measurement and assessment of performance outcomes from the various instructional strategies embedded into the games or simulations that involve the learning outcome of problem solving. Problem solving is one of the cognitive demands in the CRESST model of learning. "Problem solving is cognitive processing directed at achieving a goal when no solution method is obvious to the problem solver" (Mayer & Wittrock, 1996, p. 47). O'Neil's (1999) problem-solving model includes the following components: content understanding; problem-solving strategies—domain-independent (-general) and domain-dependent (-specific); and self-regulation, which is comprised of metacognition and motivation. Metacognition is further

comprised of self-checking/monitoring and planning, and motivation is comprised of effort and self-efficacy.

Effective problem solving in games can place a large cognitive load on working memory. Thus, instructional strategies have been recommended to help control or reduce cognitive load, for example, scaffolding (Mayer et al., 2002), worked examples (Sweller, 1988, 2003; Sweller, van Merrienboer, & Paas, 1998), and job aids. With respect to scaffolding, while there are a number of definitions (e.g., Chalmers, 2003; van Merrienboer, Clark, & deCroock, 2002; van Merrienboer, Kirschner, & Kester, 2003), what they all have in common is that scaffolding is an instructional method that provides support during learning by reducing cognitive load. Clark (2001) described instructional methods as external representations of the internal processes of selecting, organizing, and integrating. These processes provide learning goals, monitoring procedures, feedback, selection methods, hints, prompts, and various advance organizers (Alessi, 2000; Clark, 2001; Jones, Farquhar, & Surry, 1995; Leemkuil et al., 2003). Each of these components either reflects a form of scaffolding or reflects a need for scaffolding

One form of scaffolding is graphical scaffolding. A number of studies have reported the benefits of maps, which is a type of graphical scaffolding (Benbasat & Todd, 1993: Chou & Lin, 1998; Chou, Lin, & Sun, 2000; Farrell & Moore, 2000; Ruddle, Payne, & Jones, 1999). In virtual environments, navigation maps help the user to navigate, orient the user, and facilitate an easier learning experience (Yair, Mintz, & Litvak, 2001). While navigation maps can reduce or distribute cognitive load (Cobb, 1997), they also have the potential to add load, ultimately counteracting their positive effects. Navigation maps can provide valuable cognitive support for navigating virtual environments, such as computer-based video games. This can be particularly useful when using the gaming environment to accomplish a complex problem-solving task. We

currently have game research being conducted in our lab to investigate the use of both maps and worked examples.

References

Alessi, S. M. (2000). Simulation design for training and assessment. In H. F. O'Neil, Jr. & D. H. Andrews (Eds.), *Aircrew training and assessment* (pp. 197-222). Mahwah, NJ: Lawrence Erlbaum Associates.

Arthur, W., Jr., Strong, M. H., Jordan, J. A., Williamson, J. E., Shebilske, W. L., & Regian, J. W. (1995). Visual attention: Individual differences in training and predicting complex task performance. *Acta Psychologica, 88,* 3-23.

Baker, E. L., Aschbacher, P., & Bradley, C. (1985). *Evaluation of WEST: ICAI system* (Report to Jet Propulsion Laboratory). Los Angeles: University of California, Center for the Study of Evaluation.

Baker, E. L., & Mayer, R. E. (1999). Computer-based assessment of problem solving. *Computers in Human Behavior, 15,* 269-282.

Baker, E. L., & O'Neil, H. F., Jr. (2002). Measuring problem solving in computer environments: current and future states. *Computers in Human Behavior, 18,* 609-622.

BBN/DARWARS. (2004, December). *Conference presentation folder* provided at the 2004 Interservice/Industry Training, Simulation and Education Conference, Orlando, FL.

Benbasat, I., & Todd, P. (1993). An experimental investigation of interface design alternatives: Icon vs. text and direct manipulation vs. menus. *International Journal of Man-Machine Studies, 38,* 369-402.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct measurement of cognitive load in multimedia learning. *Educational Psychologist 38*(1), 53-61.

Carr, P. D., & Groves, G. (1998). The Internet-based operations simulation game. In J. A. Chambers (Ed.), *Selected papers for the 9th International Conference on College Teaching and Learning* (pp. 15-23). Jacksonville: Florida Community College at Jacksonville.

Chalmers, P. A. (2003). The role of cognitive theory in human-computer interface. *Computers in Human Behavior, 19,* 593-607.

Chou, C., & Lin, H. (1998). The effect of navigation map types and cognitive styles on learners' performance in a computer-networked hypertext learning system [Electronic Version]. *Journal of Educational Multimedia and Hypermedia, 7,* 151-176.

Chou, C., Lin, H, & Sun, C.-t. (2000). Navigation maps in hierarchical-structured hypertext courseware [Electronic Version]. *International Journal of Instructional Media, 27(2),* 165-182.

Chuang, S. (2004). *The role of search strategies and feedback on a computer-based collaborative problem-solving task.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Clark, R. E. (Ed.). (2001). *Learning from media: Arguments, analysis, and evidence.* Greenwich, CT: Information Age Publishing.

Cobb, T. (1997). Cognitive efficiency: Toward a revised theory of media. *Educational Technology Research and Development, 45*(4), 21-35.

Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86,* 1022-1033.

de Jong, T., & van Joolingen, W. R. (1998). Scientific discovery learning with computer simulations of conceptual domains. *Review of Educational Research, 68,* 179-202.

Donchin, E. (1989). The learning strategies project. *Acta Psychologica, 71,* 1-15.

Dugdale, S. (1998). Mathematical problem-solving and computers: A study of learner-initiated application of technology in a general problem-solving context. *Journal of Research on Computing in Education, 30,* 239-253.

Farrell, I. H., & Moore, D. M. (2000). The effect of navigation tools on learners' achievement and attitude in a hypermedia environment. *Journal of Educational Technology Systems, 29(2),* 169-181.

Galimberti, C., Ignazi, S., Vercesi, P., & Riva, G. (2001). Communication and cooperation in networked environment: An experimental analysis. *CyberPsychology & Behavior, 4*(1), 131-146.

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, motivation, and learning: A research and practice model. *Simulation & Gaming, 33,* 441-467.

Gopher, D., Weil, M., & Bareket, T. (1994). Transfer of skill from a computer game trainer to flight. *Human Factors, 36,* 387-405.

Gredler, M. E. (1996). Educational games and simulations: A technology in search of a (research) paradigm. In D. Jonassen (Ed.), *Handbook of research for educational communications and technology* (pp. 521-540). New York: Simon & Schuster Macmillan.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature, 423,* 534-537.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Herl, H. E., O'Neil, H. F., Jr., Chung, G., & Schacter, J. (1999) Reliability and validity of a computer-based knowledge mapping system to measure content understanding. *Computers in Human Behavior, 15,* 315-333.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research, 70,* 151-179.

Hong, J., & Liu, M. (2003). A study on thinking strategy between experts and novices of computer games. *Computers in Human Behavior, 19,* 425-458.

Hsieh, I. (2001). *Types of feedback in a computer-based collaborative problem-solving Group Task.* Unpublished doctoral dissertation, University of Southern California, Los Angeles.

Jones, M. G., Farquhar, J. D., & Surry, D. W. (1995). Using metacognitive theories to design user interfaces for computer-based learning. *Educational Technology, 35*(4), 12-22.

Kalyuga, S., Chandler, P., Touvinen, J., & Sweller, J. (2001). When problem-solving is superior to worked examples. *Journal of Educational Psychology, 93,* 579-588.

Katz, S., & Lesgold, A. (1991). Modeling the student in SHERLOCK II. In J. Kay & A. Quilici (Eds.), Proceedings of the IJCAI-91 Workshop W.4: *Agent modelling for intelligent interaction* (pp. 93-127). Sydney, Australia.

King, K. W., & Morrison, M. (1998). A media buying simulation game using the Internet. *Journalism and Mass Communication Educator, 53(3),* 28-36.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels.* San Francisco, CA: Berrett-Koehler Publishers, Inc.

Lee, J. (1999). Effectiveness of computer-based instructional simulation: A meta analysis. *International Journal of Instructional Media, 26*(1), 71-85.

Leemkuil, H., de Jong, T., de Hoog, R., & Christoph, N. (2003). KM Quest: A collaborative Internet-based simulation game. *Simulation & Gaming, 34,* 89-111.

Malone, T. W. (1981). Toward a theory of intrinsically motivating instruction. *Cognitive Science, 4,* 333-369.

Malone, T. W., & Lepper, M. R. (1987). *Aptitude, learning and instruction III: Cognitive and affective process analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Marsh, H. W., Hau, K-T., Artelt, C., & Baumert, J. (2004). *OECD's brief self-report measures of educational psychology's most useful constructs: cross-cultural, psychometric comparisons across 25 countries.* Unpublished manuscript, SELF Research Centre, University of West Sydney, Penrith South DC, New South Wales, Australia.

Mayer, R. E. (2002). A taxonomy for computer-based assessment of problem-solving. *Computers in Human Behavior, 18,* 623-632.

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? *American Psychologist, 59,* 14-19.

Mayer, R. E., Dow, G. T., & Mayer, S. (2003). Multimedia learning in an interactive self-explaining environment: What works in the design of agent-based microworlds? *Journal of Educational Psychology, 95,* 806-813.

Mayer, R. E., Mautone, P., & Prothero, W. (2002). Pictorial aids for learning by doing in a multimedia geology simulation game. *Journal of Educational Psychology, 94,* 171-185.

Mayer, R. E., & Moreno, R. (1998). A split-attention effect in multimedia learning: evidence for dual processing systems in working memory. *Journal of Educational Psychology, 90,* 312-320.

Mayer, R. E., Sobko, K., & Mautone, P. D. (2003). Social cues in multimedia learning: Role of speaker's voice. *Journal of Educational Psychology, 95,* 419-425.

Mayer, R. E., & Wittrock, M. C. (1996). Problem-solving transfer. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 47-62). New York: Simon & Schuster Macmillan.

Mislevy, R. J. (in press). Substance and structure in assessment arguments. *Law, Probability, and Risk.* preprint at http://www.education.umd.edu/EDMS/mislevy/papers/Argument.pdf>

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1,* 3-67.

Mislevy, R. J., Wilson, M. R., Ercikan, K., & Chudowsky, N. (2003). Psychometric principles in student assessment. In T. Kellaghan & D. Stufflebeam (Eds.), *International handbook of educational evaluation* (pp. 489-531). Dordrecht, the Netherlands: Kluwer Academic Press.

Moreno, R., & Mayer, R. E. (2000). Engaging students in active learning: The case for personalized multimedia messages. *Journal of Educational Psychology, 92,* 724-733.

O'Neil, H. F., Jr. (1999). Perspectives on computer-based performance assessment of problem solving: Editor's introduction. *Computers in Human Behavior, 15,* 255-268.

O'Neil, H. F., Jr. (2002). Perspective on computer-based assessment of problem solving [Special Issue]. *Computers in Human Behavior, 18,* 605-607.

O'Neil, H. F., Jr., & Fisher, Y.-C. (2004). A technology to support leader development: Computer games. In D. V. Day, S. J. Zaccaro, & S. M. Halpin (Eds.), *Leader development for transforming organizations* (pp. 99-121). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Neil, H. F., Jr., & Herl, H. E. (1998, April). *Reliability and validity of a trait measure of self-regulation.* Presented at the annual meeting of the American Educational Research Association, San Diego, CA.

O'Neil, H. F., Jr., Baker, E. L., & Fisher, J. Y.-C. (2002). *A formative evaluation of ICT games* (Report to the Institute for Creative Technologies, University of Southern California). Los Angeles: University of Southern California, Rossier School of Education.

Parchman, S. W., Ellis, J. A., Christinaz, D., & Vogel, M. (2000). An evaluation of three computer-based instructional strategies in basic electricity and electronics training. *Military Psychology, 12*(1), 73-87

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist, 37,* 91-105.

Porter, D. B., Bird, M. E., & Wunder, A. (1990-1991). Competition, cooperation, satisfaction, and the performance of complex tasks among Air Force cadets. *Current Psychology: Research & Reviews, 9,* 347-354.

Prislin, R., Jordan, J. A., Worchel, S., Semmer, F. T., & Shebilske, W. L. (1996, September). Effects of group discussion on acquisition of complex skills. *Human Factors, 38,* 404-416.

Quinn, C. N. (1991). Computers for cognitive research: A HyperCard adventure game. *Behavior Research Methods, Instruments, & Computers, 23,* 237-246.

Ramsberger, P. F., Hopwood, D., Hargan, C. S., & Underhill, W. G. (1983), *Evaluation of a spatial data management system for basic skills education. Final Phase I Report for Period 7 October 1980 - 30 April 1983* (HumRRO FR-PRD-83-23). Alexandria, VA: Human Resources Research Organization.

Rhodenizer, L., Bowers, C. A., & Bergondy, M. (1998). Team practice schedules: What do we know? *Perceptual and Motor Skills, 87,* 31-34.

Ricci, K. E., Salas, E., & Cannon-Bowers, J. A. (1996). Do computer-based games facilitate knowledge acquisition and retention? *Military Psychology, 8,* 295-307.

Rieber, L. P., & Parmley, M. W. (1995). To teach or not to teach? Comparing the use of computer-based simulations in deductive versus inductive approaches to learning with adults in science. *Journal of Educational Computing Research, 13,* 359-374.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130,* 261-288.

Rosenorn, T., & Kofoed, L. B. (1998). Reflection in learning processes through simulation/gaming. *Simulation & Gaming, 29,* 432-440.

Ruben, B. D. (1999). Simulations, games, and experience-based learning: The quest for a new paradigm for teaching and learning. *Simulation & Gaming, 30,* 498-505.

Ruddle, R. A., Payne, S. J., & Jones, D. M. (1999). The effects of maps on navigation and search strategies in very-large-scale virtual environments. *Journal of Experimental Psychology: Applies, 5(1),* 54-75.

Schacter, J., Herl, H. E., Chung, G., Dennis, R. A., & O'Neil, H. F., Jr. (1999). Computer-based performance assessments: A solution to the narrow measurement and reporting of problem-solving. *Computers in Human Behavior, 13,* 403-418.

Schank, R. C. (1997). *Virtual learning: A revolutionary approach to build a highly skilled workforce.* New York: McGraw-Hill Trade.

Seifert, T. L. (2004). Understanding student motivation. *Educational Research, 46,* 137-149.

Shebilske, W. L., Regian, W., Arthur, W., Jr., & Jordan, J. A. (1992). A dyadic protocol for training complex skills. *Human Factors, 34,* 369-374.

Sugrue, B., & Kim, K.-H. (2004). *State of the industry. ASTD's annual review of trends in workplace learning and performance. Executive summary.* Alexandria, VA: American Society for Training and Development.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12,* 257-285.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction, 4,* 295-312.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning and motivation* (Vol. 43, pp. 215-266). San Diego: Academic Press.

Sweller, J., Van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251-269.

Thiagarajan, S. (1998). The myths and realities of simulations in performance technology. *Educational Technology, 38*(4), 35-41.

Thomas, P., & Macredie, R. (1994). Games and the design of human-computer interfaces. *Educational Technology, 31,* 134-142.

Tkacz, S. (1998). Learning map interpretation: Skill acquisition and underlying abilities. *Journal of Environmental Psychology, 18,* 237-249.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge: Cambridge University Press.

van Merrienboer, J. J. G., Clark, R. E., & de Croock, M. B. M. (2002). Blueprints for comp lex learning: The 4C/ID-model. *Educational Technology Research & Development, 50(2),* 39-64.

van Merrienboer, J. J. G., Kirschner, P. A., & Kester, L. (2003). Taking a load off a learner's mind: Instructional design for complex learning. *Educational Psychologist, 38*(1), 5-13.

Wainess R., & O'Neil, H. F., Jr. (2003). *Feasibility study.* Unpublished manuscript, University of Southern California, Rossier School of Education.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16,* 3-118.

Wolfe, J. (1997, December). The effectiveness of business games in strategic management course work [Electronic Version]. Simulation & Gaming Special Issue: *Teaching Strategic Management, 28,* 360-376.

Yair, Y., Mintz, R., & Litvak, S. (2001). 3D-virtual reality in science education: An implication for astronomy teaching. *Journal of Computers in Mathematics and Science Teaching, 20,* 293-305.